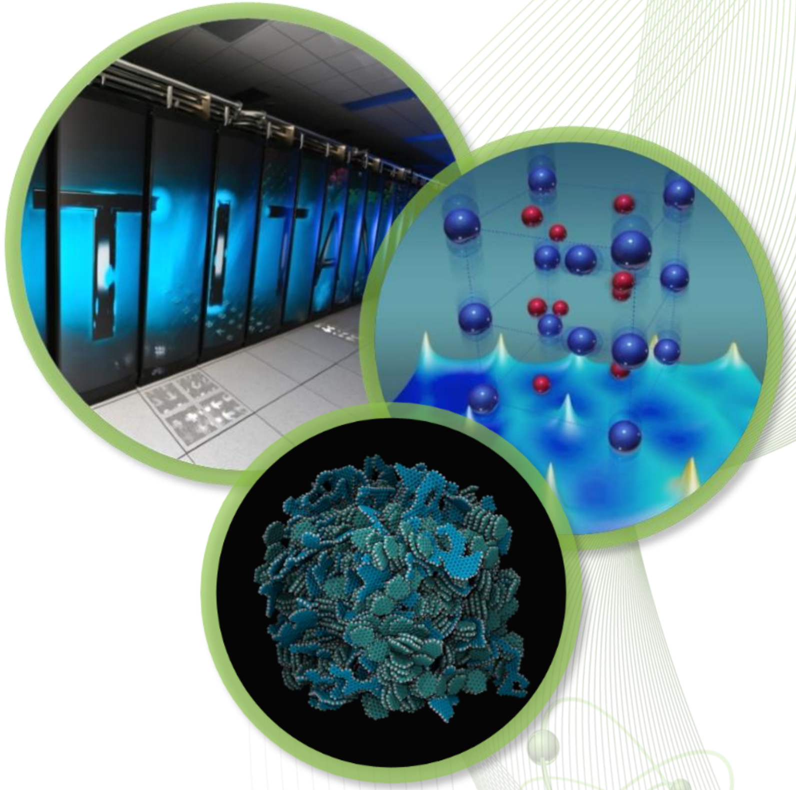
Population Level Deep Learning: Scalable Information Extraction From Clinical Pathology Reports with CANDLE



Arvind Ramanathan, Jacob Hinkle, Fernanda Foertter (on behalf of Pilot 3/CANDLE)

Computational Science & Engineering, Biomedical Sciences, Engineering and Computing (BSEC), Health Data Sciences Institute (HDSI), Oak Ridge National Laboratory (ORNL)

http://ramanathanlab.org

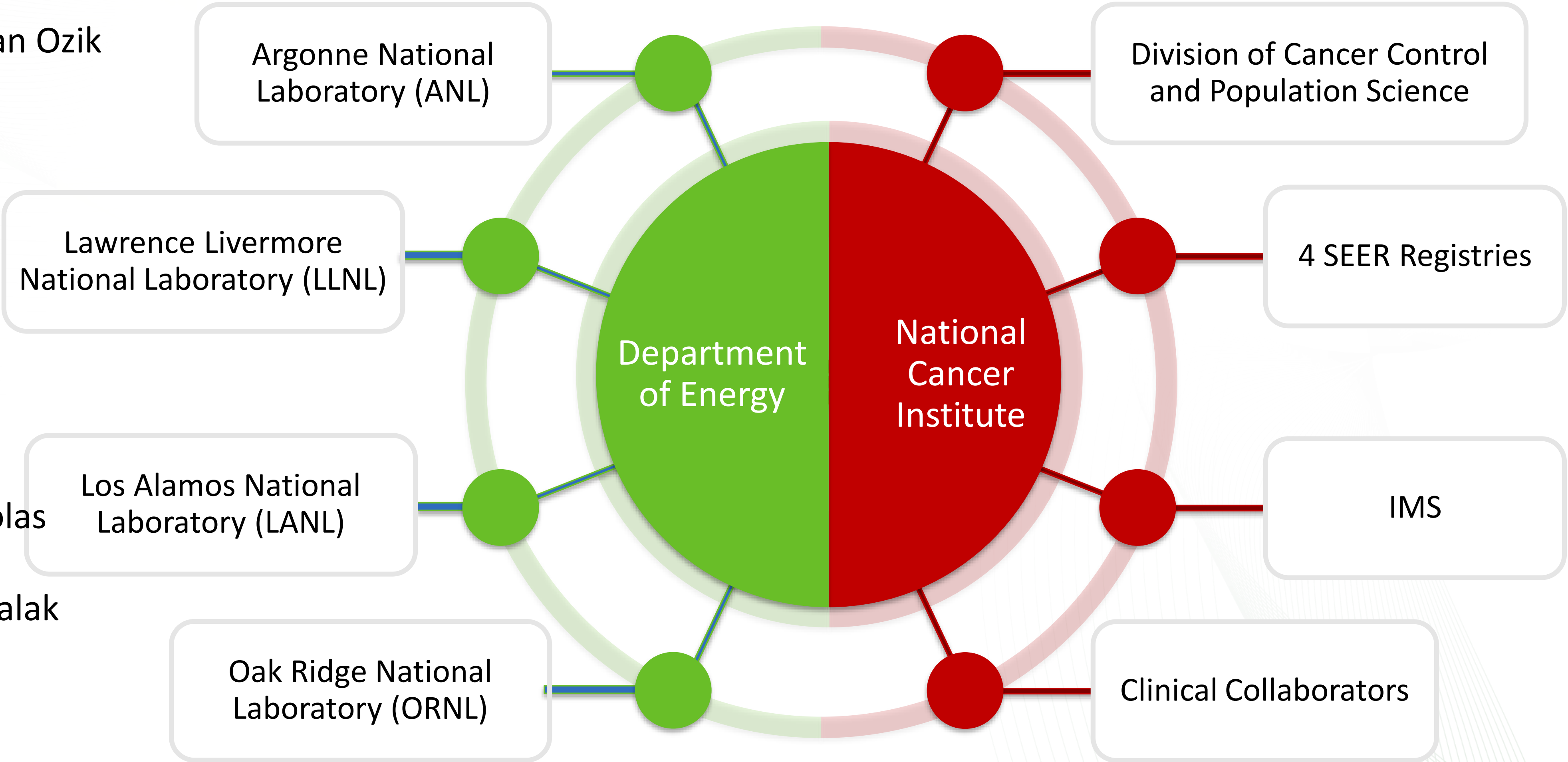
ramanathana@ornl.gov

Team

Tom Bretin, Jonathan Ozik

Ana Paula De Oliveira Sales, Priyadip Ray, Braden Soper

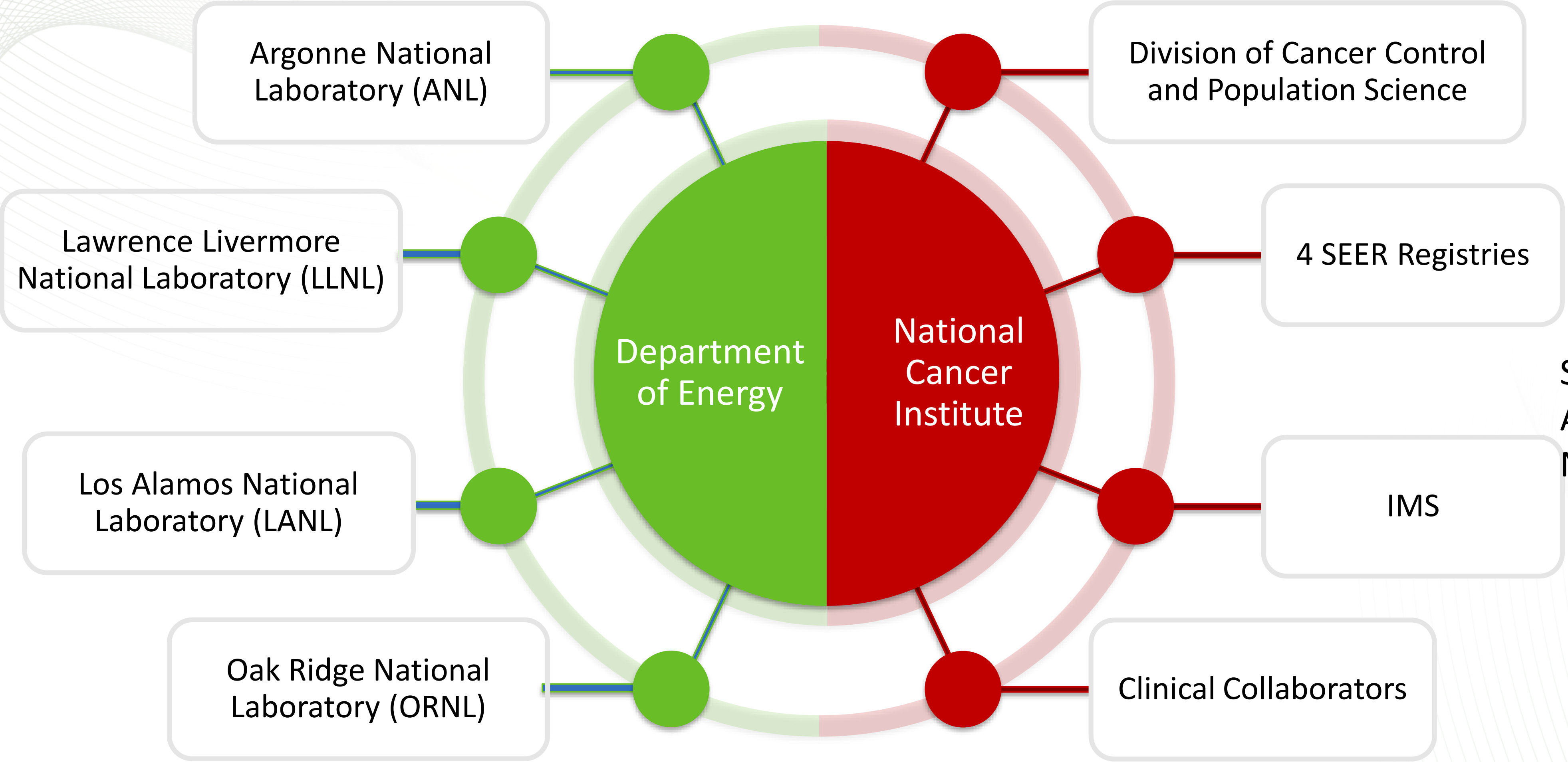
Tanmoy Bhattacharya, Kumkum Ganguly, Nicholas Hengartner, Benjamin MacMahon, Sarah Michalak



Folami Alamudun Mohammed Alawad, Blair Christian, Shang Gao, John Qiu, Kshitij Shrivastava, Georgia Tourassi, Hong-Jun Yoon, Todd Young

Team

Jessica Boten, Paul Fearn, Rocky Feuer, Usman Khalid, Marina Matatova, Spencer Morris, Mita Myneni, Lynne Penberthy

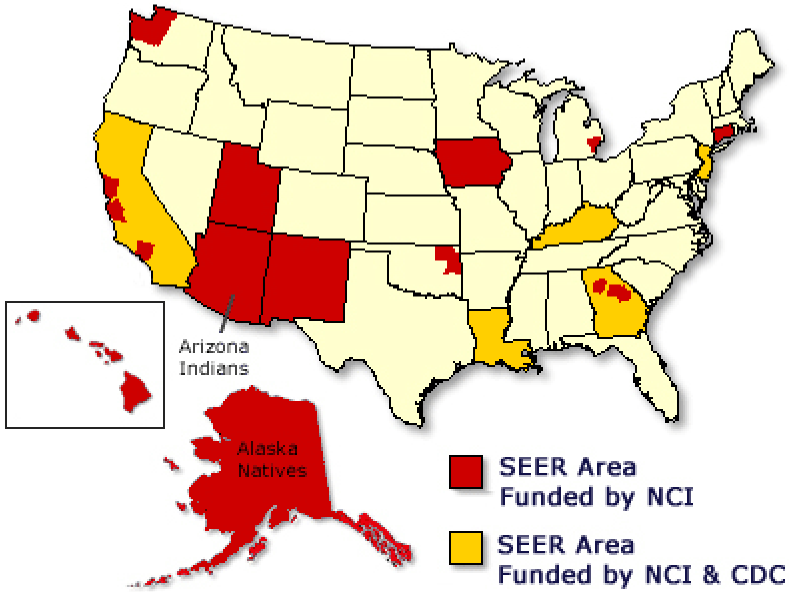


LA, KY, Seattle, GA

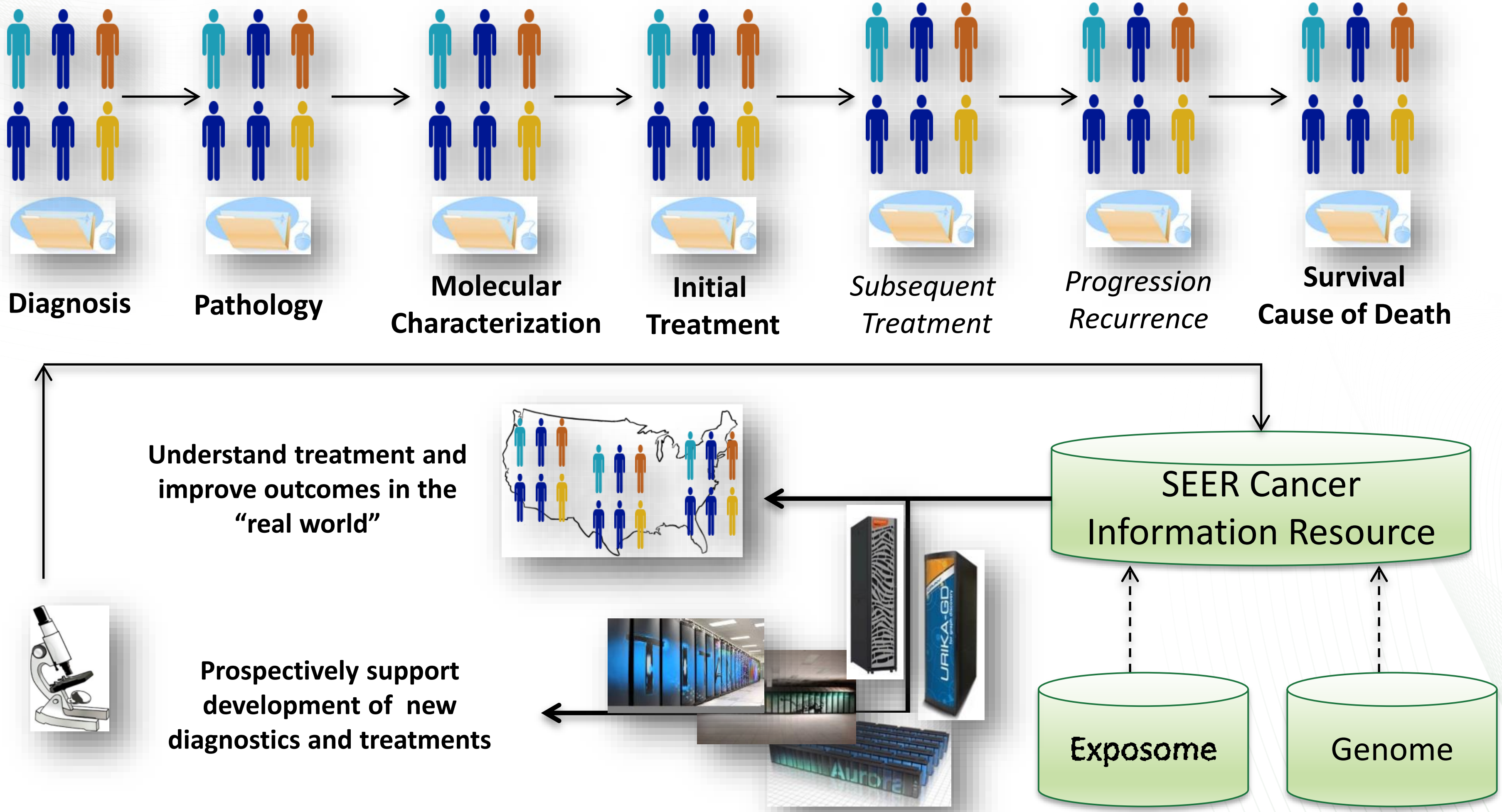
Scott Depoy, Linda Coyle, Dav
Annett
Nicola Schussler

SEER Program Overview

- Funded by NCI ***to support research*** on the diagnosis, treatment and outcomes of cancer since 1973
- Population-based registries covering ~28% of the US population
  - Representing racial and ethnic minorities
  - Various geographic subgroups
- 450,000+ incident cases reported annually
  - Approximately 85% of cases with real time electronic pathology reporting
  - Collect survival and cause of death outcomes
- Impact (1973-2016)
  - >4500 downloads per year
  - 7398 publications using SEER data for analysis
  - 40,230 publications referencing SEER data
  - >191,000 SEER*Stat users annually
  - Study planning, recruitment, and follow-up
  - Annual Report to the Nation on the Status of Cancer



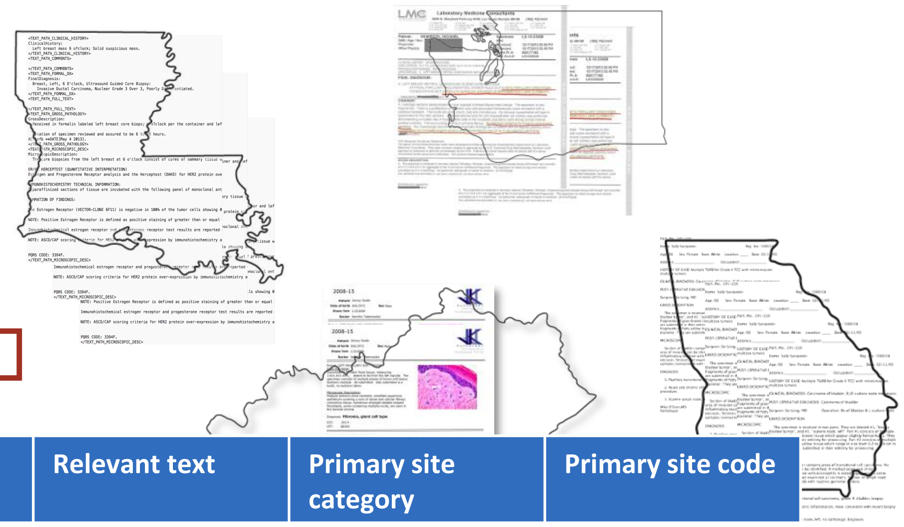
Cancer Surveillance Pilot: Improve the effectiveness of cancer treatment in the “real world” through computing



Cancer Pathology Report Processing Pipeline

NCI SEER Database

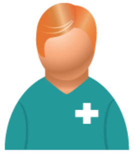
Integration with structured data from Electronic medical records for patients



Registry	PatientID	Record No.	Tumor No.	Primary Site	Source Section	Relevant text	Primary site category	Primary site code
KY	114431		3	Breast	Final diagnosis	Mammary carcinoma	Breast	C50.9 Breast,NOS
KY	118420		5	Breast	Final diagnosis	BREAST PRIMARY	BREAST	C50.9 Breast, NOS
SE	0084621	500713999	01	Lung	Final diagnosis	Lung, right lower lobe	lung	C34.3 lower lobe, lung

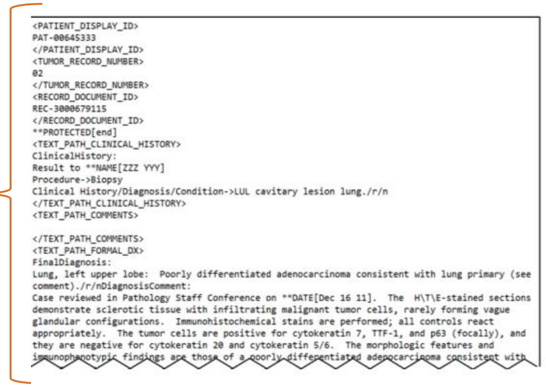
ion and

Patient

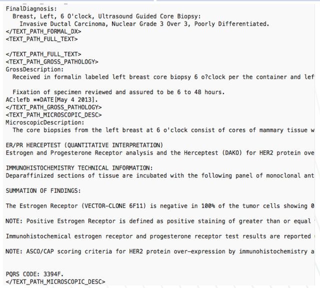


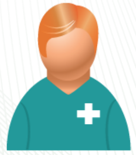
Pathologist

Diagnosis by a pathologist analyzing tissue specimen from patient



```
<PATIENT_DISPLAY_ID>
PAT-00645333
</PATIENT_DISPLAY_ID>
<TUMOR_RECORD_NUMBER>
02
</TUMOR_RECORD_NUMBER>
<RECORD_DOCUMENT_ID>
REC-3000679115
</RECORD_DOCUMENT_ID>
**PROTECTED[end]
<TEXT_PATH_CLINICAL_HISTORY>
ClinicalHistory:
Result to **NAME[ZZZ YYY]
Procedure->Biopsy
Clinical History/Diagnosis/Condition->LUL cavitary lesion lung./r/n
</TEXT_PATH_CLINICAL_HISTORY>
<TEXT_PATH_COMMENTS>

</TEXT_PATH_COMMENTS>
<TEXT_PATH_FORMAL_DX>
FinalDiagnosis:
Lung, left upper lobe:  Poorly differentiated adenocarcinoma consistent with lung primary (see
comment)./r/nDiagnosisComment:
Case reviewed in Pathology Staff Conference on **DATE[Dec 16 11].  The  H\T\E-stained sections
demonstrate sclerotic tissue with infiltrating malignant tumor cells, rarely forming vague
glandular configurations.  Immunohistochemical stains are performed; all controls react
appropriately.  The tumor cells are positive for cytokeratin 7, TTF-1, and p63 (focally), and
they are negative for cytokeratin 20 and cytokeratin 5/6.  The morphologic features and
immunophenotypic findings are those of a poorly differentiated adenocarcinoma consistent with
```





Certified Tumor Registrar

CTR at a cancer registry reviews complete patient medical record + path report

NCI-SEER is a primary data source… need to modernize

- **NEED**
  - Abstracting structured data from free-text pathology reports is critical for the national cancer surveillance program
- **CHALLENGE**
  - Manual abstraction is time-consuming, costly, and not scalable
- **GOAL**
  - Develop a scalable framework for automated information extraction from pathology reports


```
<TEXT_PATH_CLINICAL_HISTORY>
ClinicalHistory:
  Left breast mass 6 o?clock; Solid suspicious mass.
</TEXT_PATH_CLINICAL_HISTORY>
<TEXT_PATH_COMMENTS>

</TEXT_PATH_COMMENTS>
<TEXT_PATH_FORMAL_DX>
FinalDiagnosis:
  Breast, Left, 6 O'clock, Ultrasound Guided Core Biopsy:
    Invasive Ductal Carcinoma, Nuclear Grade 3 Over 3, Poorly Differentiated.
</TEXT_PATH_FORMAL_DX>
<TEXT_PATH_FULL_TEXT>

</TEXT_PATH_FULL_TEXT>
<TEXT_PATH_GROSS_PATHOLOGY>
GrossDescription:
  Received in formalin labeled left breast core biopsy 6 o?clock per the container and lef

  Fixation of specimen reviewed and assured to be 6 to 48 hours.
AC:lefb **DATE[May 4 2013].
</TEXT_PATH_GROSS_PATHOLOGY>
<TEXT_PATH_MICROSCOPIC_DESC>
MicroscopicDescription:
  The core biopsies from the left breast at 6 o'clock consist of cores of mammary tissue w

ER/PR HERCEPTEST (QUANTITATIVE INTERPRETATION)
Estrogen and Progesterone Receptor analysis and the Herceptest (DAKO) for HER2 protein ove

IMMUNOHISTOCHEMISTRY TECHNICAL INFORMATION:
Deparaffinized sections of tissue are incubated with the following panel of monoclonal ant

SUMMATION OF FINDINGS:

The Estrogen Receptor (VECTOR-CLONE 6F11) is negative in 100% of the tumor cells showing 0

NOTE: Positive Estrogen Receptor is defined as positive staining of greater than or equal

Immunohistochemical estrogen receptor and progesterone receptor test results are reported

NOTE: ASCO/CAP scoring criteria for HER2 protein over-expression by immunohistochemistry a


PQRS CODE: 3394F.
</TEXT_PATH_MICROSCOPIC_DESC>
```

Specific Aims

Deep Text Comprehension for information capture

Advanced machine learning for scalable patient Information capture from unstructured clinical reports to semi-automate the SEER program

Novel data analytic techniques for patient information integration

Scalable graph and visual analytics to understand the association between patient trajectories and patient outcomes

Data-driven integrated modeling and simulation for precision oncology

Precision modeling of patient trajectories

In silico clinical trials

State-of-the-Art Approaches in Clinical NLP

- **Current NLP thinking is TASK-specific**
- **Rule-based -** effective but require intense domain expert involvement
  - Task-specific dictionaries of phrases and medical terms
  - *Manual effort not easily scalable across tasks*
- **Conventional machine learning** - scalable but require intense feature engineering
  - N-gram based
  - Concept-extraction-based methods
- **Deep Learning -** scalable with enough compute power and ***enough*** *data*
  - Does not require dictionaries, not susceptible to misspellings etc.
  - Lots of new DL architectures proposed for NLP
  - No clear winner – depends on the global semantics required for the task at hand

Datasets Used for Preliminary Research

STUDY 1: Limited dataset of de-identified breast and lung cancer electronic pathology (e-path) reports from 5 different SEER registries

~2,500 breast and lung cancer de-identified e-path reports

Partially annotated for **subsite, laterality, grade, behavior**

STUDY 2: Large dataset of e-path reports from Louisiana Tumor Registry housed at the PHI enclave within ORNL

~267,000 reports from Louisiana Tumor Registry (2004-2017)

Gold standard for **site, laterality, grade, behavior, histology** derived from consolidated “Cancer/Tumor/Case” (CTC) records

Experimental Pipeline

Data Pre-processing

- Duplicate records
- Non-contradicting labels
- Incorrect organ annotations
- Small sample sizes
- Corpus curation

Feature Representation

- TF-IDF
- Bag-of-graphs
- RAKE
- CHUNK
- GLOVE

Training Protocols

With over sampling

Without over sampling

Rule-based Systems (RL)

- Contextualize (keywords for topics of interest)
- Term identification
- Classification

Machine Learning (ML)

- Naïve Bayes (NB)
- Logistic Regression (LR)
- Random Forest (RF)
- Support Vector Machines (SVM)
- Extreme gradient boosting tree (Xgboost)

Deep Learning (DL)

- Convolutional neural nets (CNN)
- Hierarchical Attention nets (HAN)
- Multi-task Deep neural net (MT-DNN)

Performance Metrics

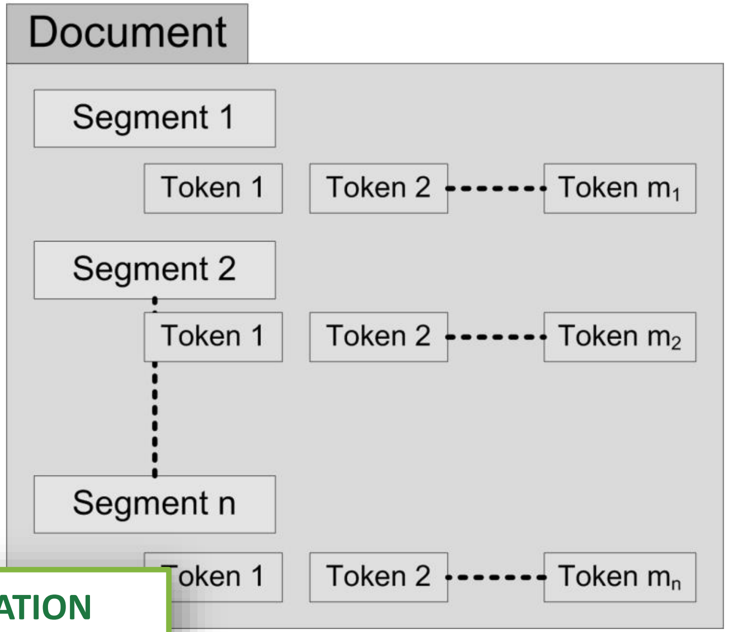
- Precision (positive predictive value) / Recall (sensitivity) / F1 per class
- Macro / Micro scores (aggregate performance over all)

Validation Strategies

- K-fold cross validation (K-fold)
- Leave-one-registry out (LORO)
- Leave-one-case-out per registry (LOO_R)

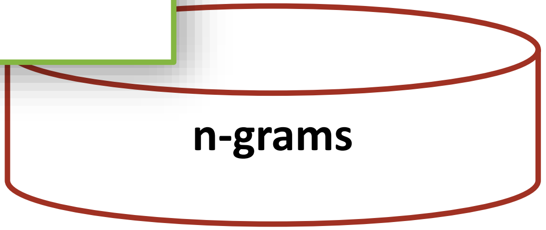
Document Representation

```
<TEXT_PATH_CLINICAL_HISTORY>
ClinicalHistory:
  Left breast mass 6 o?clock; Solid suspicious mass.
</TEXT_PATH_CLINICAL_HISTORY>
<TEXT_PATH_COMMENTS>

</TEXT_PATH_COMMENTS>
<TEXT_PATH_FORMAL_DX>
FinalDiagnosis:
  Breast, Left, 6 O'clock, Ultrasound Guided Core Biopsy:
    Invasive Ductal Carcinoma, Nuclear Grade 3 Over 3, Poorly Differentiated.
</TEXT_PATH_FORMAL_DX>
<TEXT_PATH_FULL_TEXT>

</TEXT_PATH_FULL_TEXT>
<TEXT_PATH_GROSS_PATHOLOGY>
GrossDescription:
  Received in formalin labeled left breast core biopsy 6 o?clock per the container and lef

  Fixation of specimen reviewed and assured to be 6 to 48 hours.
AC:lefb **DATE[May 4 2013].
</TEXT_PATH_GROSS_PATHOLOGY>
<TEXT_PATH_MICROSCOPIC_DESC>
MicroscopicDescription:
  The core biopsies from the left breast at 6 o'clock consist of cores of mammary tissue w

ER/PR HERCEPTEST (QUANTITATIVE INTERPRETATION)
Estrogen and Progesterone Receptor analysis a

IMMUNOHISTOCHEMISTRY TECHNICAL INFORMATION:
Deparaffinized sections of tissue are incuba

SUMMATION OF FINDINGS:

The Estrogen Receptor (VECTOR-CLONE 6F11) is

NOTE: Positive Estrogen Receptor is defined a

Immunohistochemical estrogen receptor and pr

NOTE: ASCO/CAP scoring criteria for HER2 pro


PQRS CODE: 3394F.
</TEXT_PATH_MICROSCOPIC_DESC>
```



Feature Representation

- TF-IDF
- Bag-of-graphs
- RAKE
- CHUNK
- GLOVE



Output

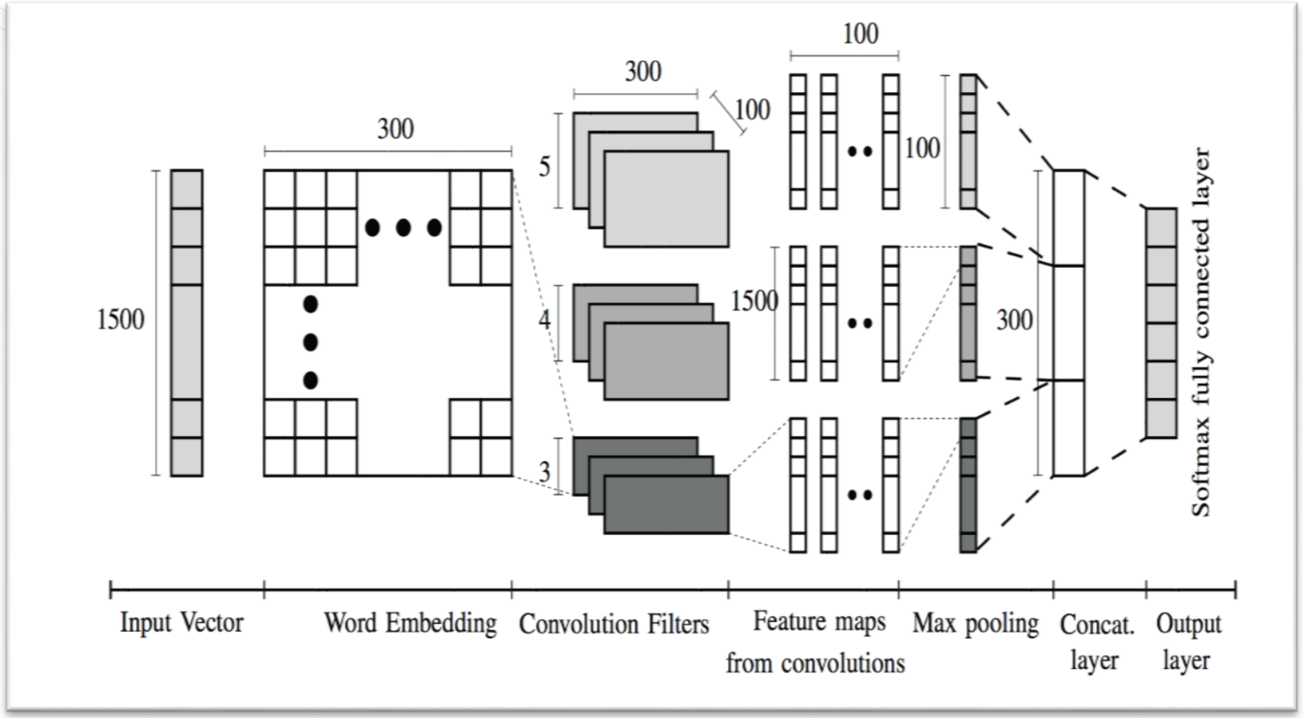
tumor	shows	small	cell	carcinoma
tumor shows	shows small	small cell	cell carcinoma	carcinoma cell

Document Representation

	t_1	t_2	…	t_n
S_1				
…				
S_p				

A ‘gentle’ introduction to convolutional nets (CNN) for text

Given a document represented as a collection of words, how do we extract features automatically?



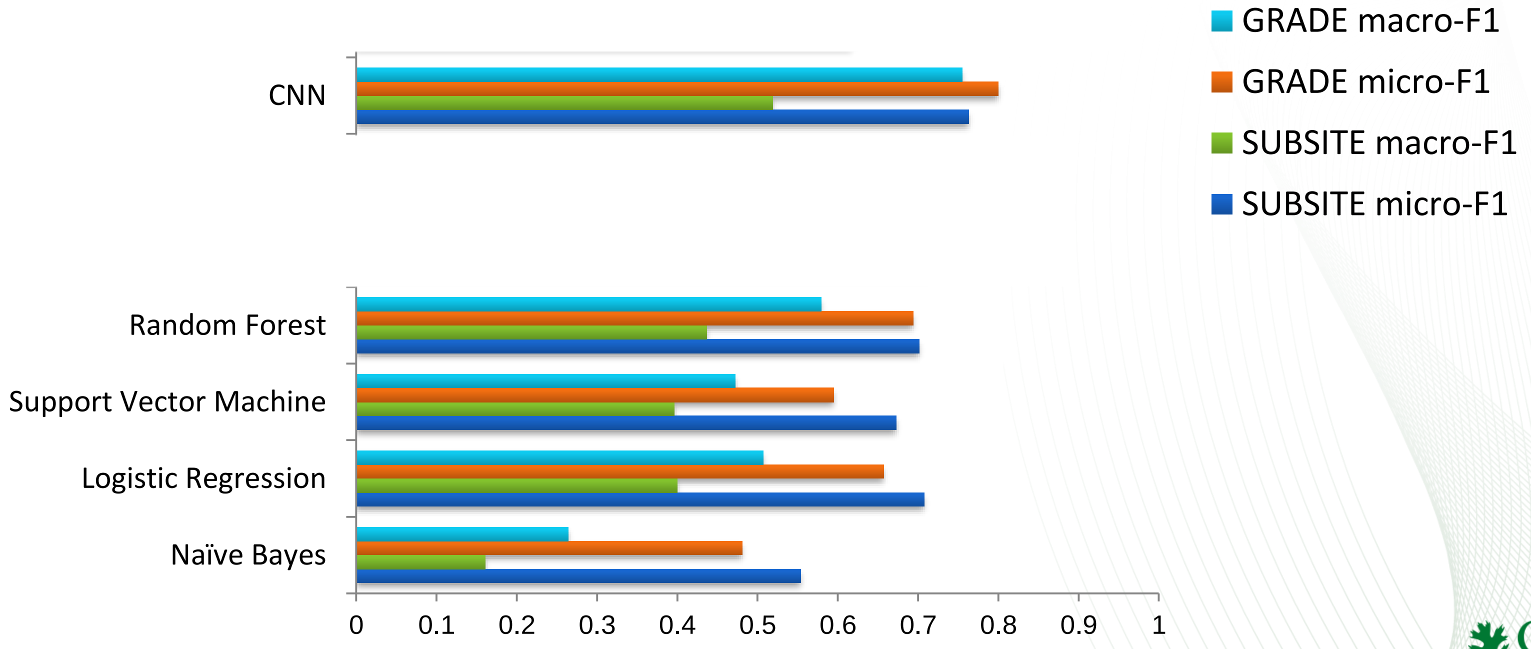
- Text is presented in the form of a document matrix – a sequence of word embedding vectors
- Multiple convolutional filters capture context along a document:
  - Word lengths {3,4,5} are used to “slide” along the entire length
- Network learns to select context features in via max pooling
- Selected features are concatenated and fed though a fully connected layer where regularization occurs
- Output is finally a softmax classifier

“Deep Learning for Automated Extraction of Primary Sites from Cancer Pathology Reports,” IEEE Journal of Biomedical and Health Informatics [January 2018]

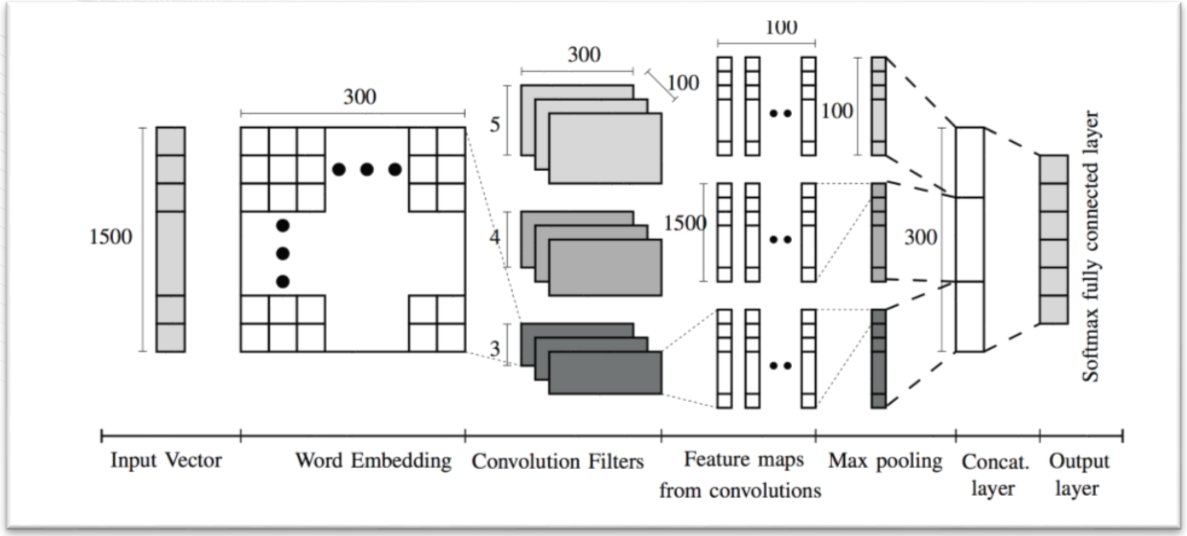
CNNs perform better in basic information extraction tasks compared to conventional ML approaches

Deep Learning

Conventional Machine Learning



CANDLE hyper-parameter optimization boosts performance



Hyper-parameter Optimization

1. Word embedding method
2. Word embedding size
3. No. of convolution filters
4. Size of convolution filters
5. No. of fully connected layers
6. Size of fully connected layers

	Primary Site		Grade	
	Micro-F	**Macro-F**	**Micro-F**	**Macro-F**
Empirical optimization (May 2017)	0.712	0.398	0.716	0.521
HyperSpace optimization (October 2017)	0.763	0.519	0.800	0.755

Highlights & Caveats of using CNN for text

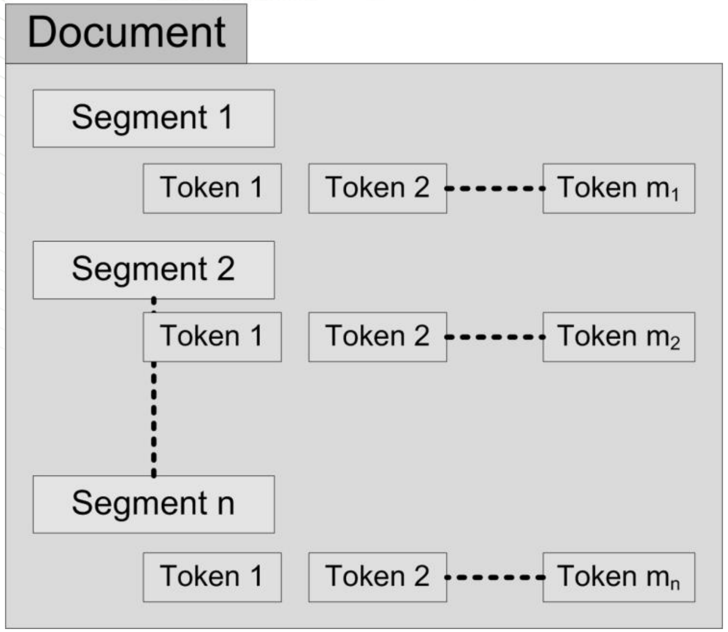
Highlights

- CNN learns features automatically:
  - Context is discerned directly from word embedding
- CNNs can abstract concepts relatively well with less user intervention
- Modifications to convolutions is relatively simple

Caveats

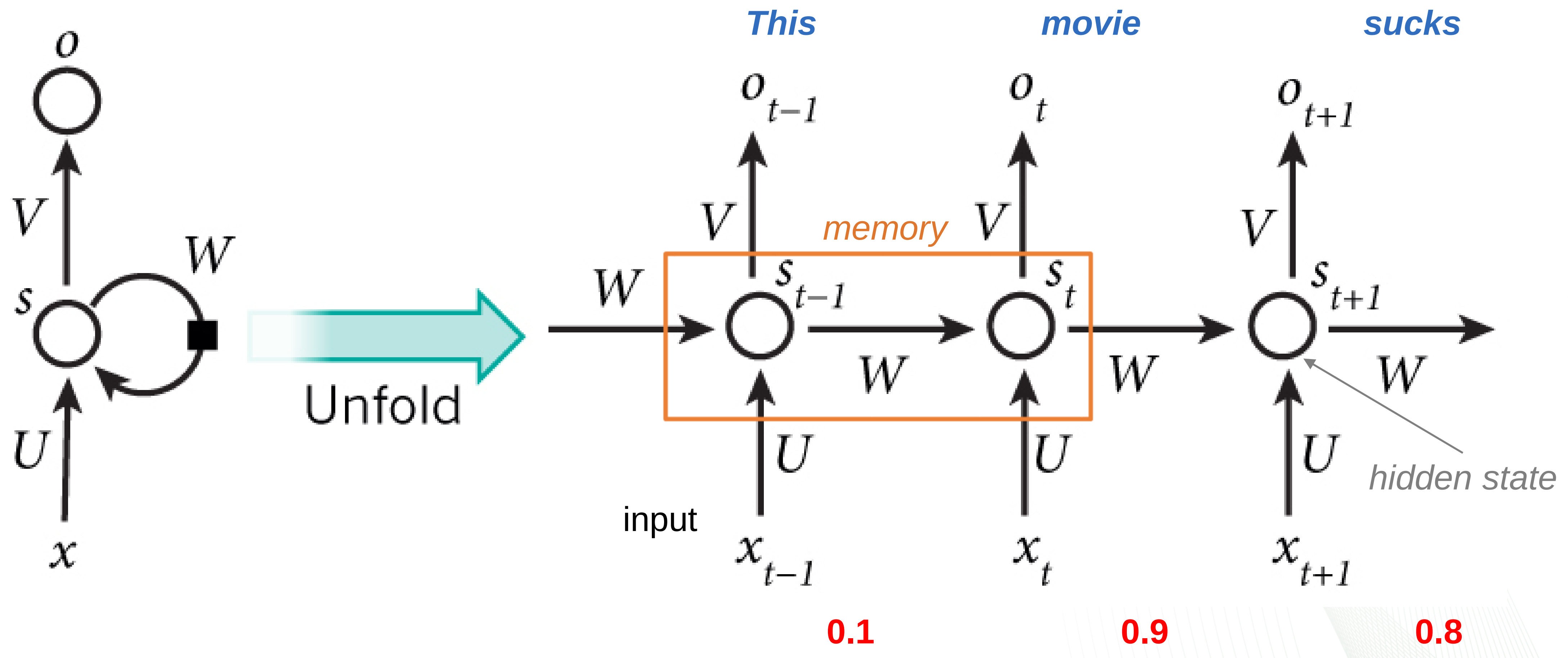
- Context extraction is sensitive:
  - ***Location variance***: where does a word occur or co-occur is important
  - ***Compositionality***: adjective modifying a noun, medical terms have specific meanings depending on what occurs before & after.
- Need larger corpus to achieve good levels of task-level performance

Building a slightly sophisticated model for documents



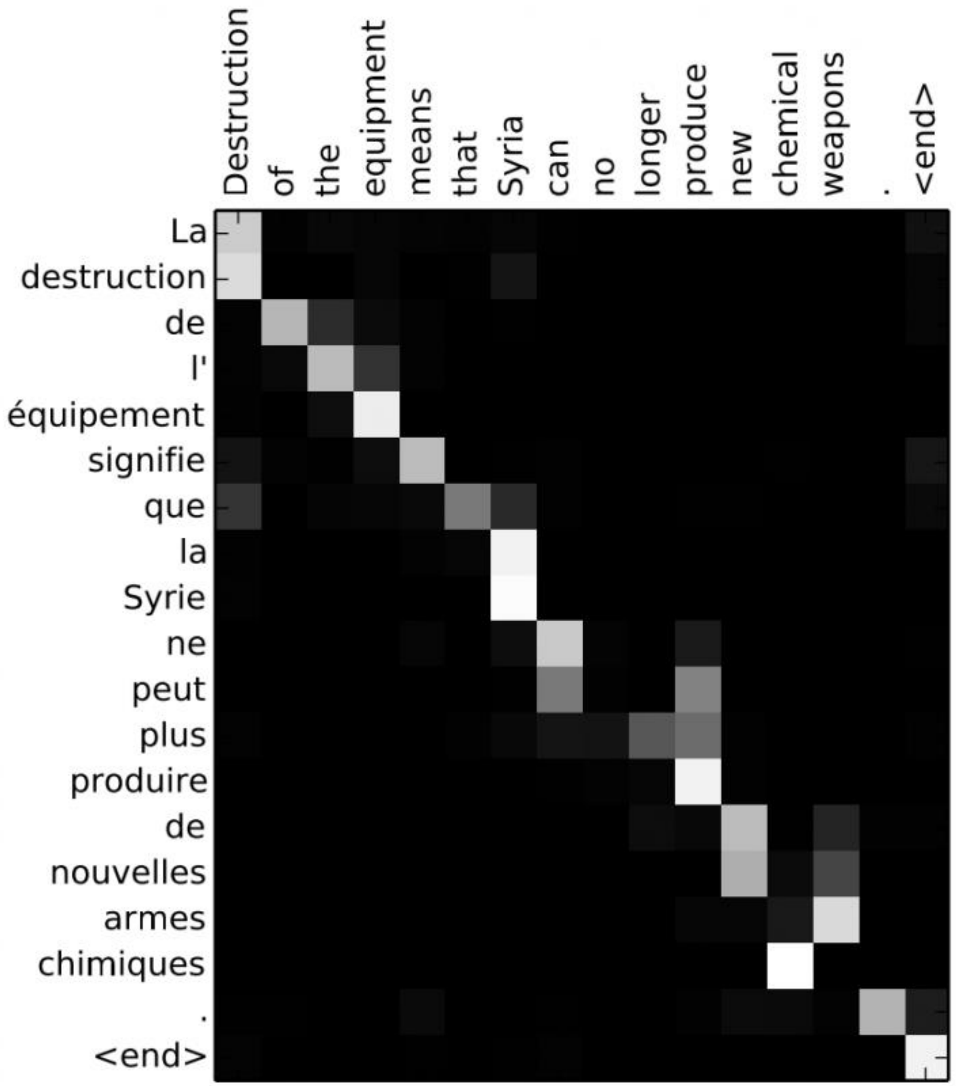
- Documents are formed of sentences read from left to right (in order)
  - Distinct sequence representation
- Probability of emitting the next word in the sequence is dependent on a "hierarchy":
  - Sentences formed of words
  - Documents formed of sentences
  - 2 level hierarchy
- Can we capture this behavior automatically?

Sequential modeling with Recurrent neural networks (RNN)



- Variety of applications: (1) Speech recognition, (2) Language translation, (3) Video prediction
- Sequence modeling takes care of location variance in sentences

Capturing context and relevance through attention mechanisms in RNN

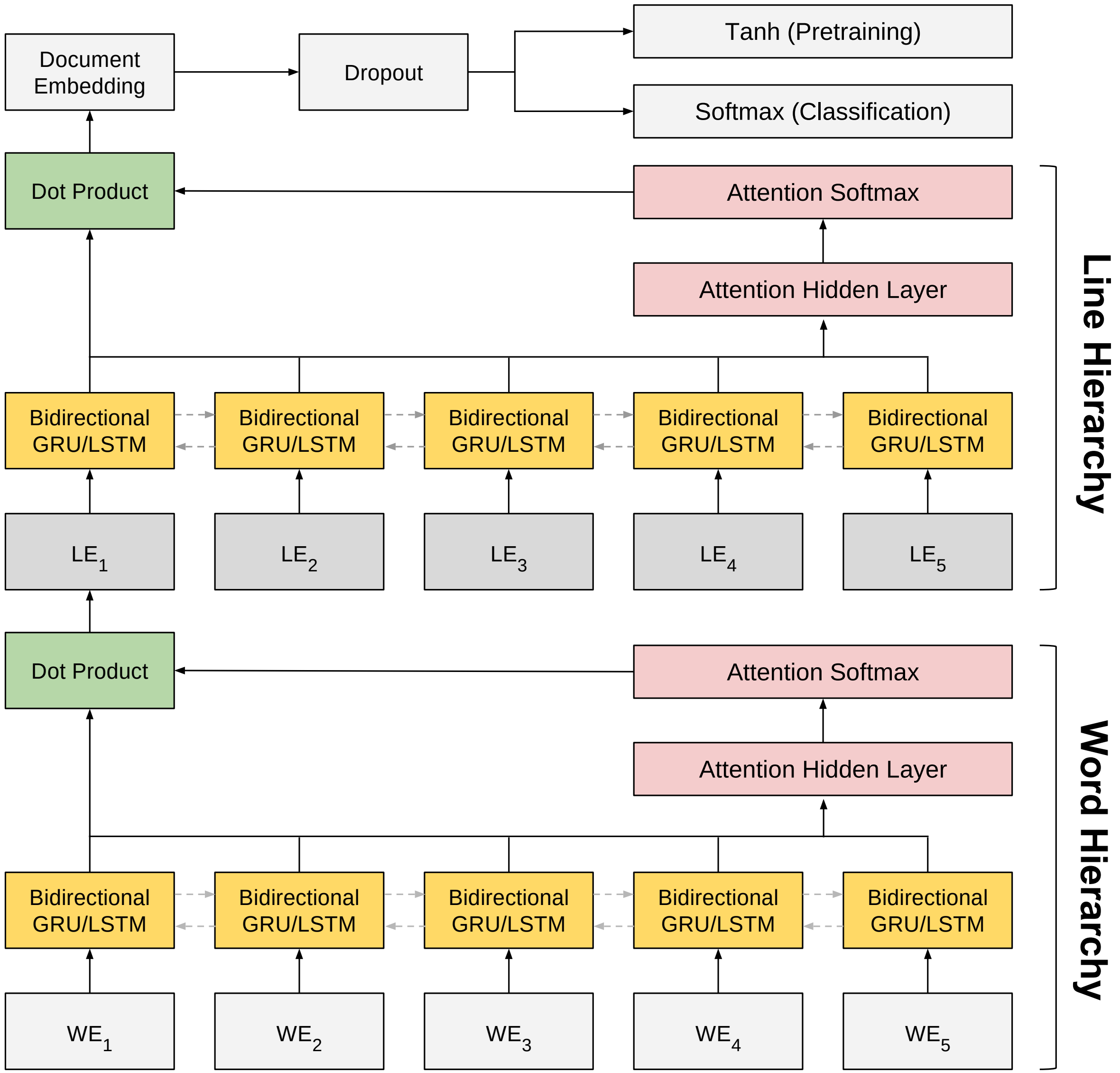


- x_t is a word in a sentence that is being generated using some underlying "sequence"
- Every y_t is produced by some "decoder" depends on a **weighted combination of all the input states**, not just the last state
- *a's* define the weights for each input state

Neural Machine Translation by Jointly Learning to Align and Translate
Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio, ICLR 2015

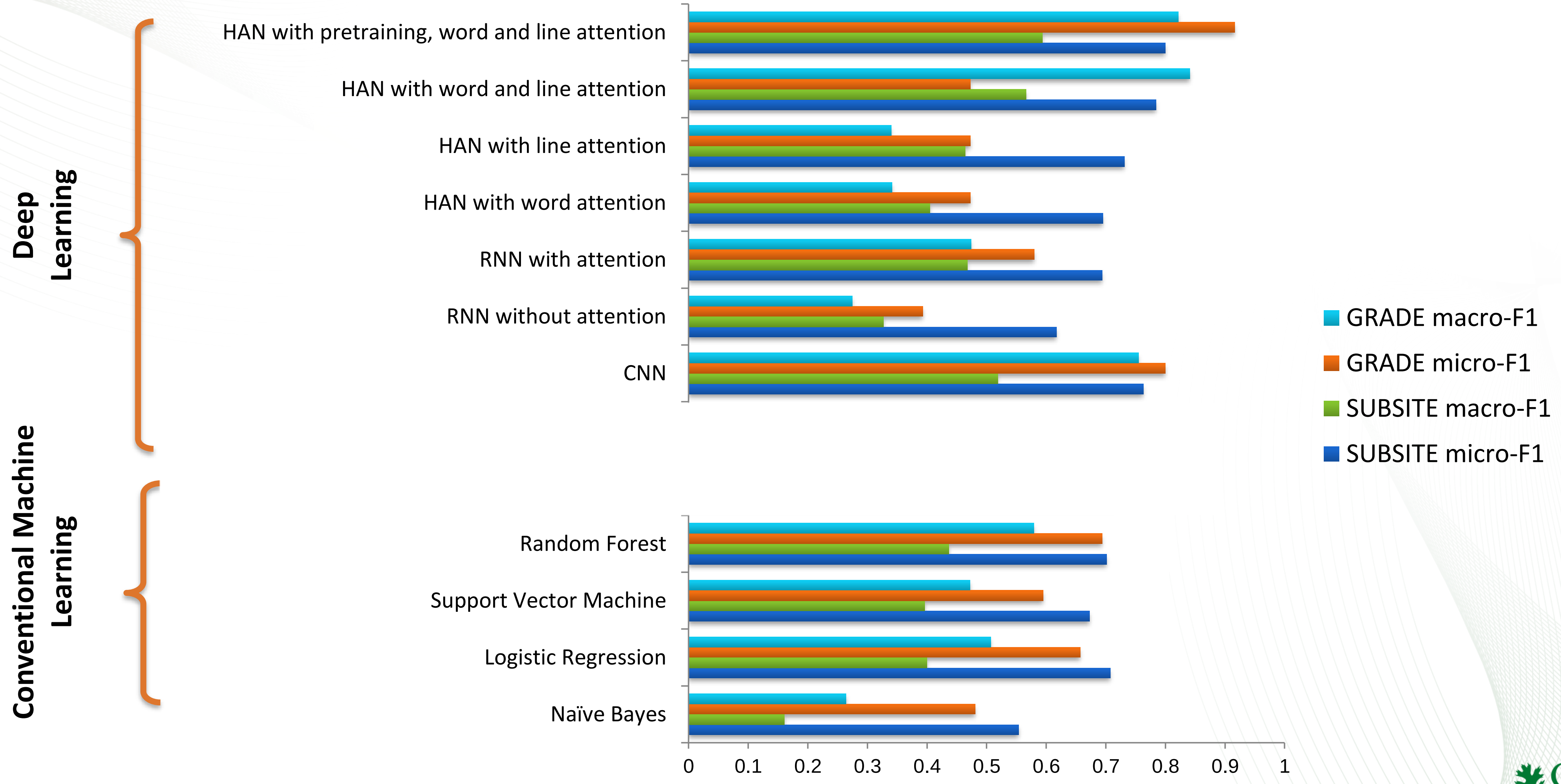
Layering an RNN with attention… Hierarchical attention network (HAN)

- Word level embedding:
  - capture important words in a sentence
  - ***Output***: sentence embedding weighted based on word occurrence/ co-occurrence most relevant for classification task
- Sentence level embedding:
  - capture important sentences within a document
  - **Output**: weighted sentence embedding based on relevance for classification task
- Final document embedding is fed into classification



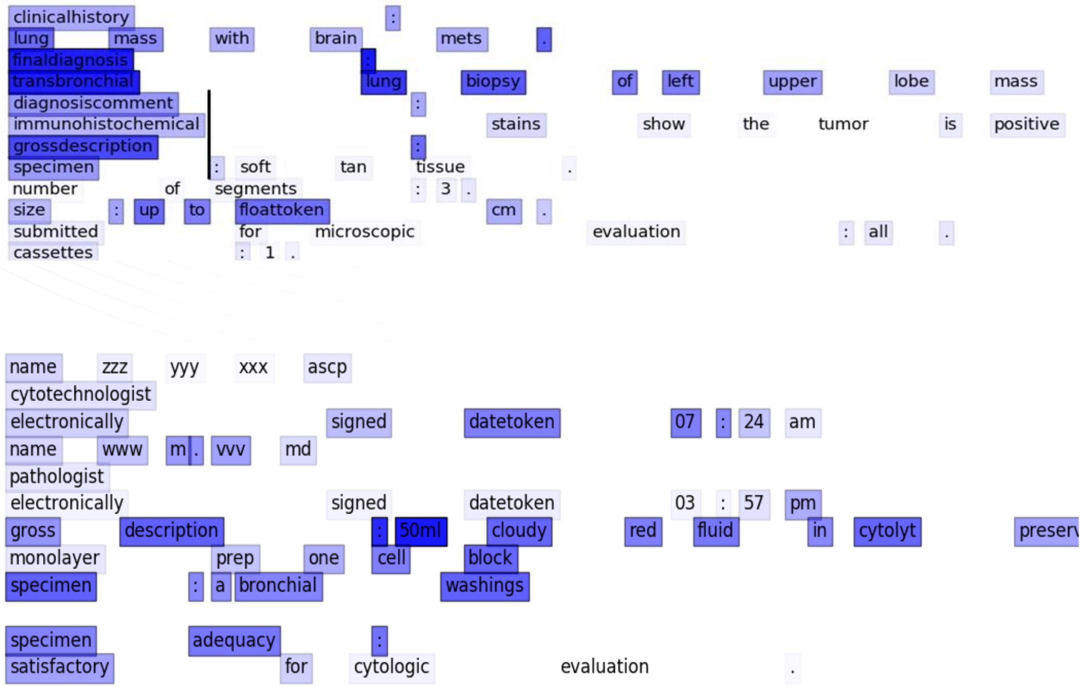
Hierarchical Attention Networks for Information Extraction from Cancer Pathology Reports," Journal of American Medical Informatics Association [appeared online, Nov 2017]

CNNs perform better in basic information extraction tasks compared to conventional ML approaches

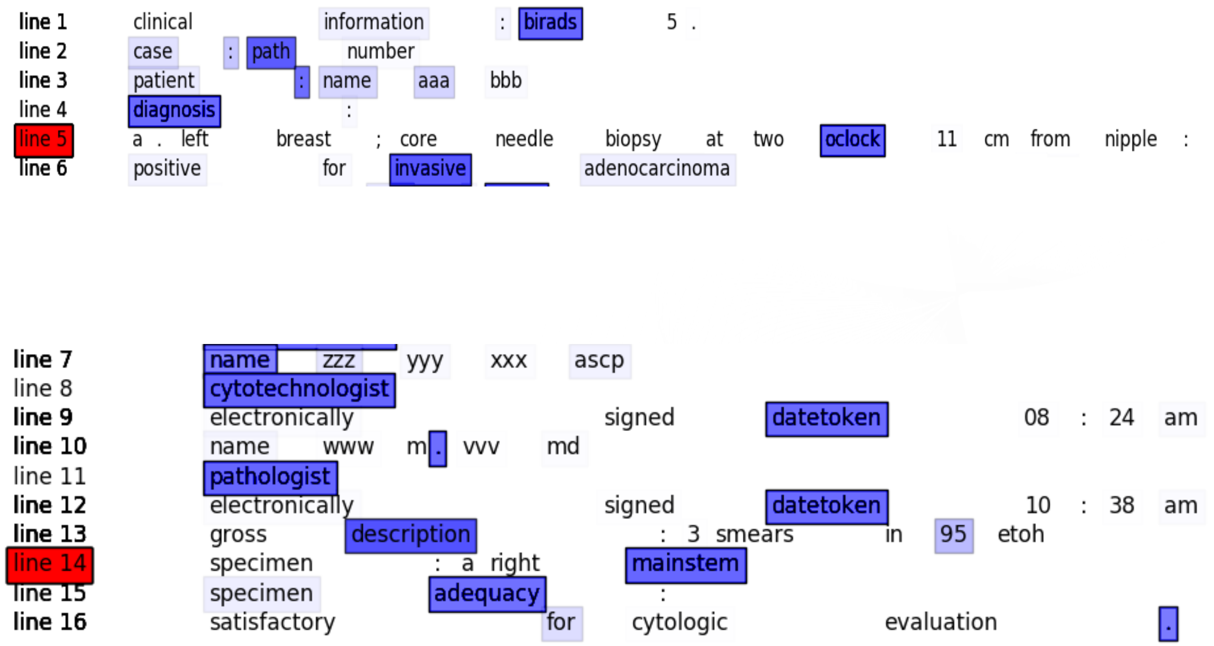


Interpreting what CNNs and HANs learned from ePath reports

CNN



HAN



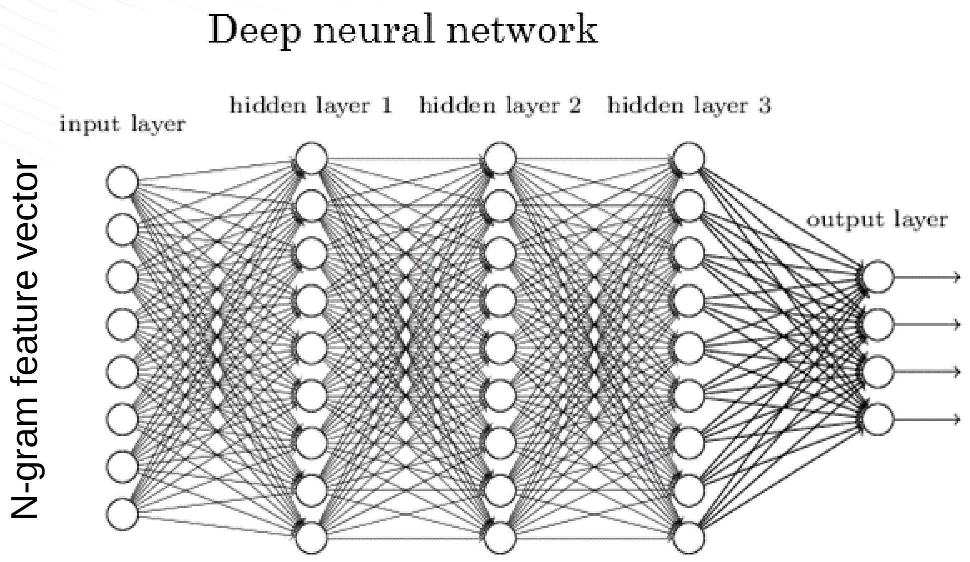
CNNs blindly associate context with importance based on how often words occur in its neighborhood. Moving along a row, these words may not always capture the required clinical context.

HANs interpret context based on most important words in a sentence → sentences → document. Neighboring words/sentences provide overall importance.

Extending the number of classification tasks

Single Task DNN

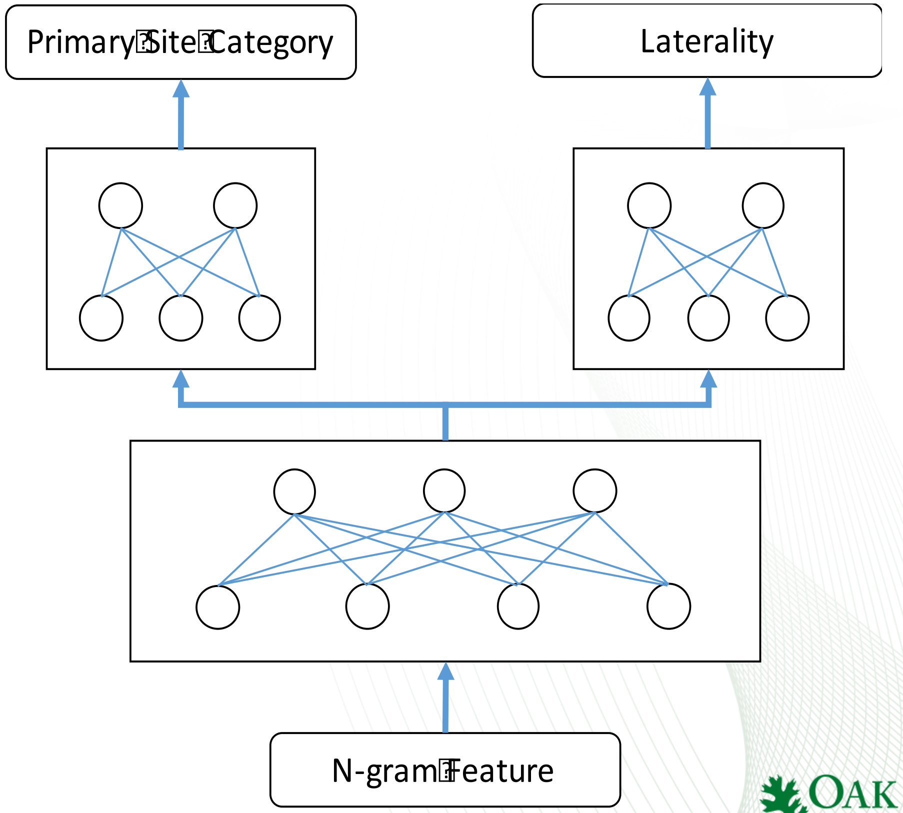
- Output layer produces class probability over k classes using the softmax nonlinearity
- Stochastic gradient descent



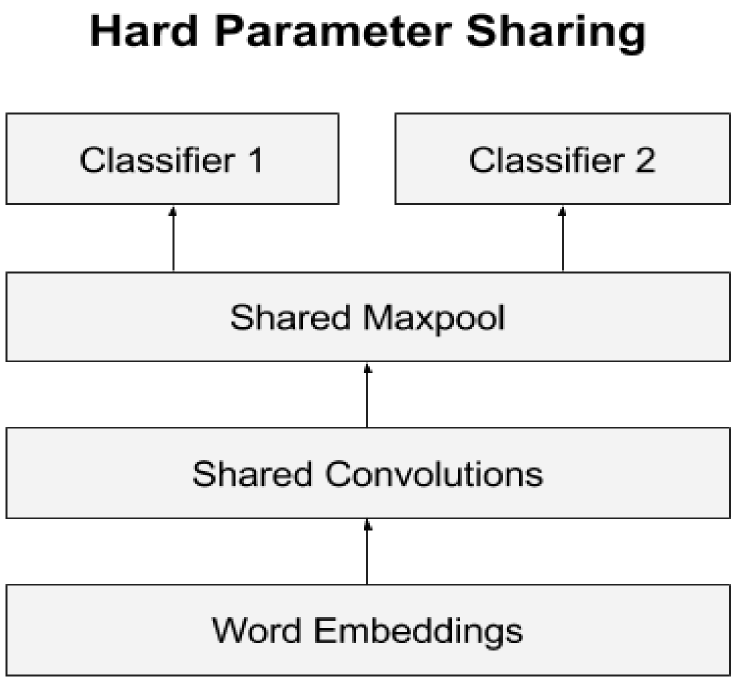
HJ Yoon, A. Ramanathan, G.D. Tourassi, "Multi-task Deep Neural Networks for Automated Extraction of Primary Site and Laterality Information from Cancer Pathology Reports." In INNS Conference on Big Data [2016]

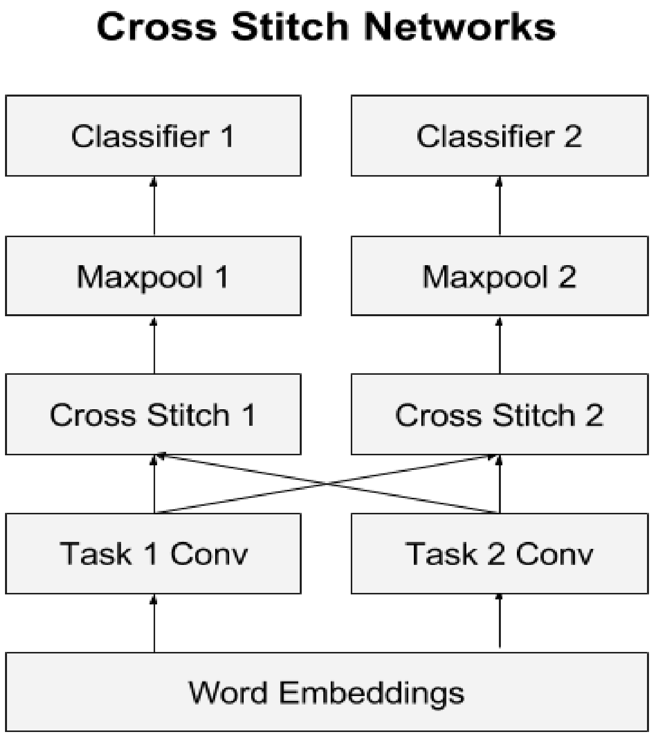
Multi-Task DNN

- Exploits tasks relatedness
- Multiple tasks solved simultaneously
- Trained with same optimization technique and document representation as singe task DNN



Multi-Task CNNs: Two different implementations



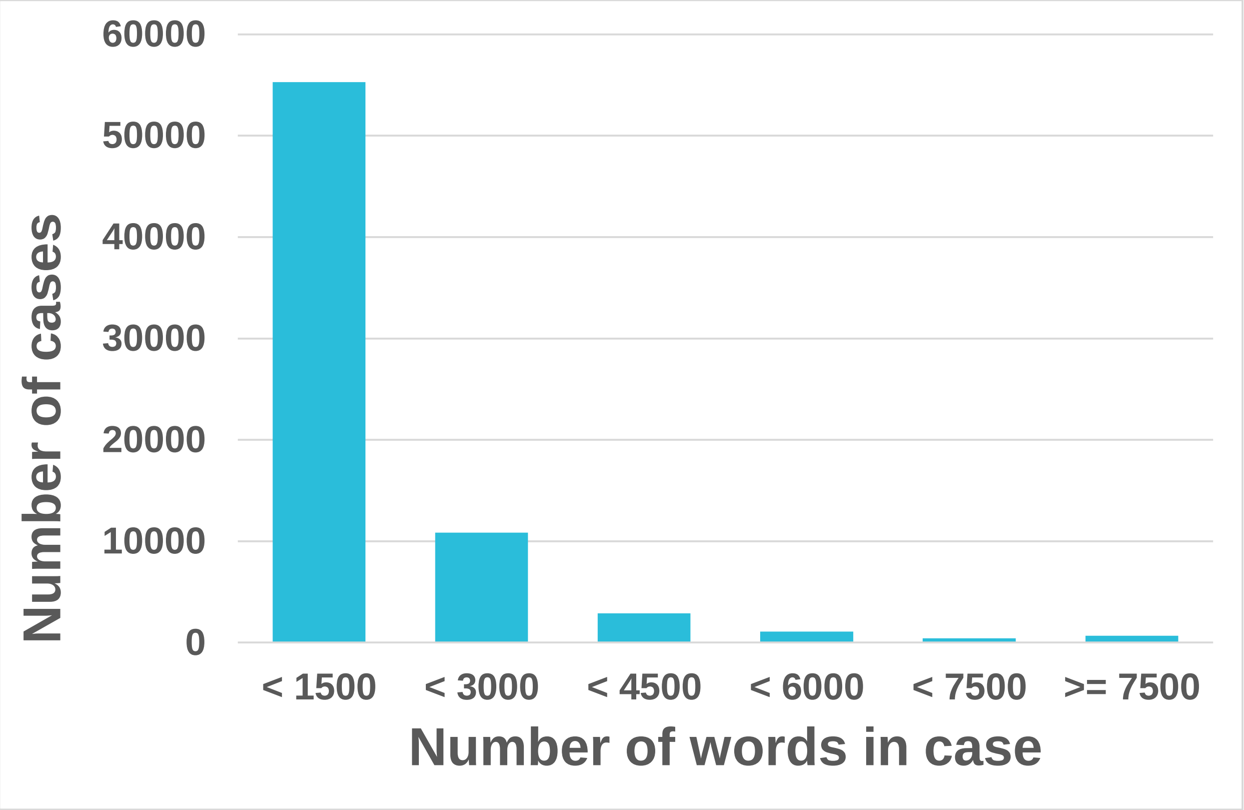


- The same convolutional layers are used for all tasks
- These convolutional layers find shared features that are useful across all tasks
- Each task has its own softmax classifier

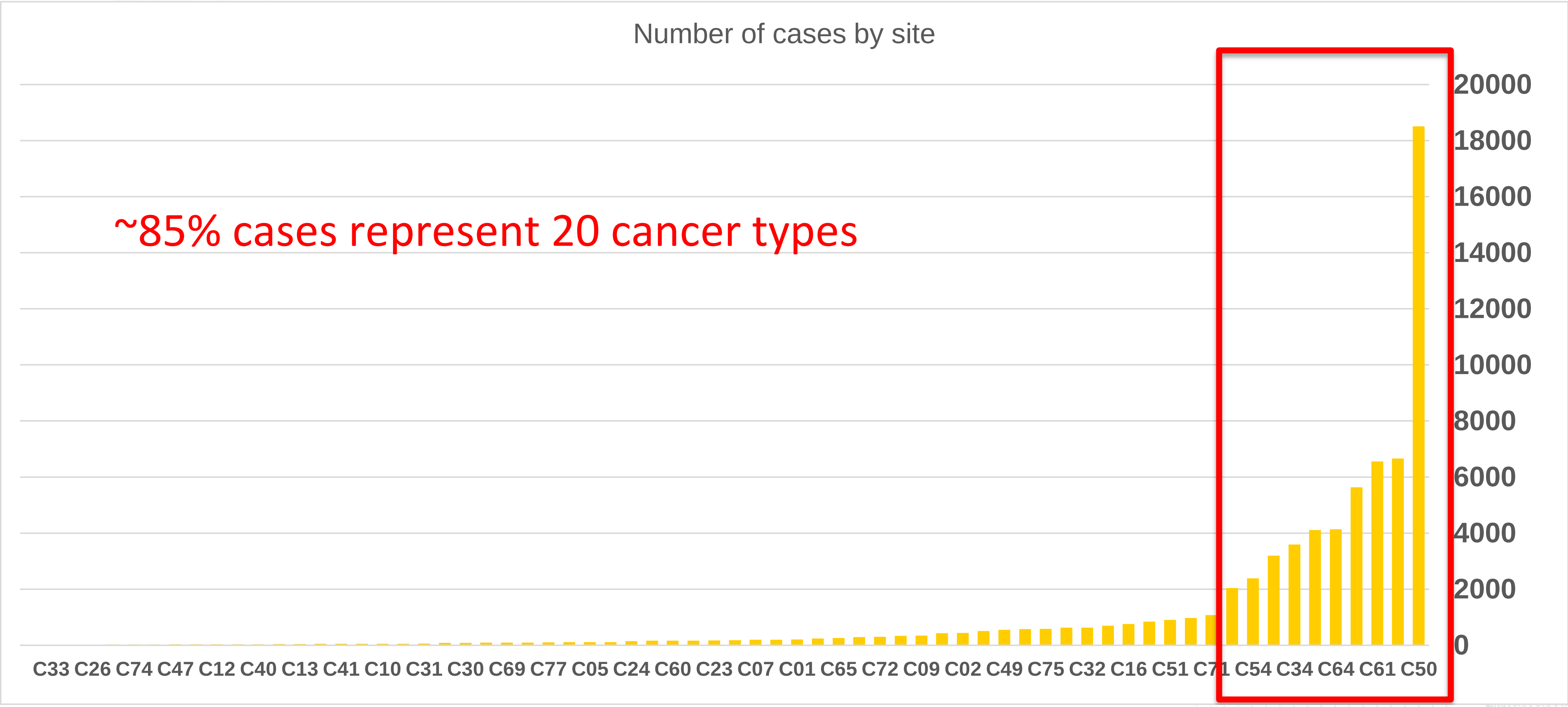
- Each task has its own set of convolutional layers
- A cross stitch operation learns the best linear combination of features from each task
- Each task has its own softmax classifier

STUDY 2: Benchmarking CNN on Louisiana Registry Path Corpus

- 2004-2017
- 71,223 tumors
- 2-fold cross-validation on 2004-2015 (59,427 cases)
- Additional testing 2016-2017 (11,796 cases)



5 information extraction tasks: Site, Histology, Laterality, Grade, Behavior)

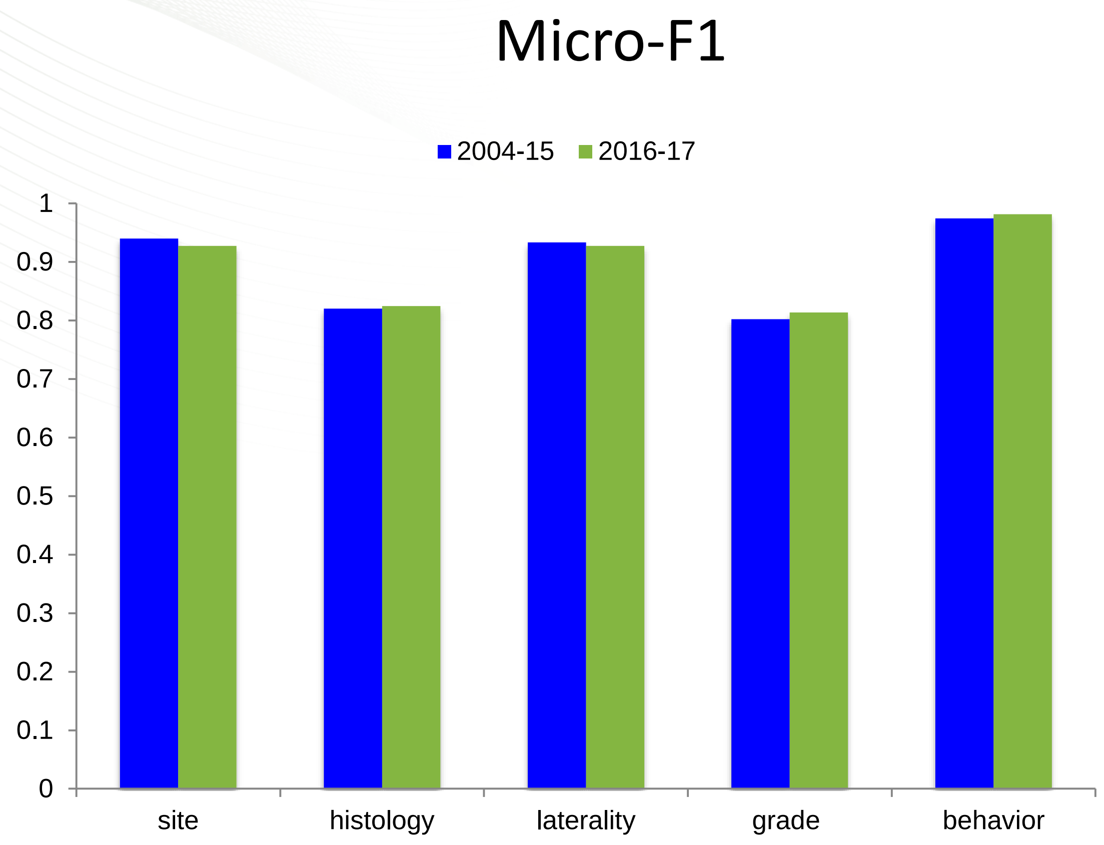


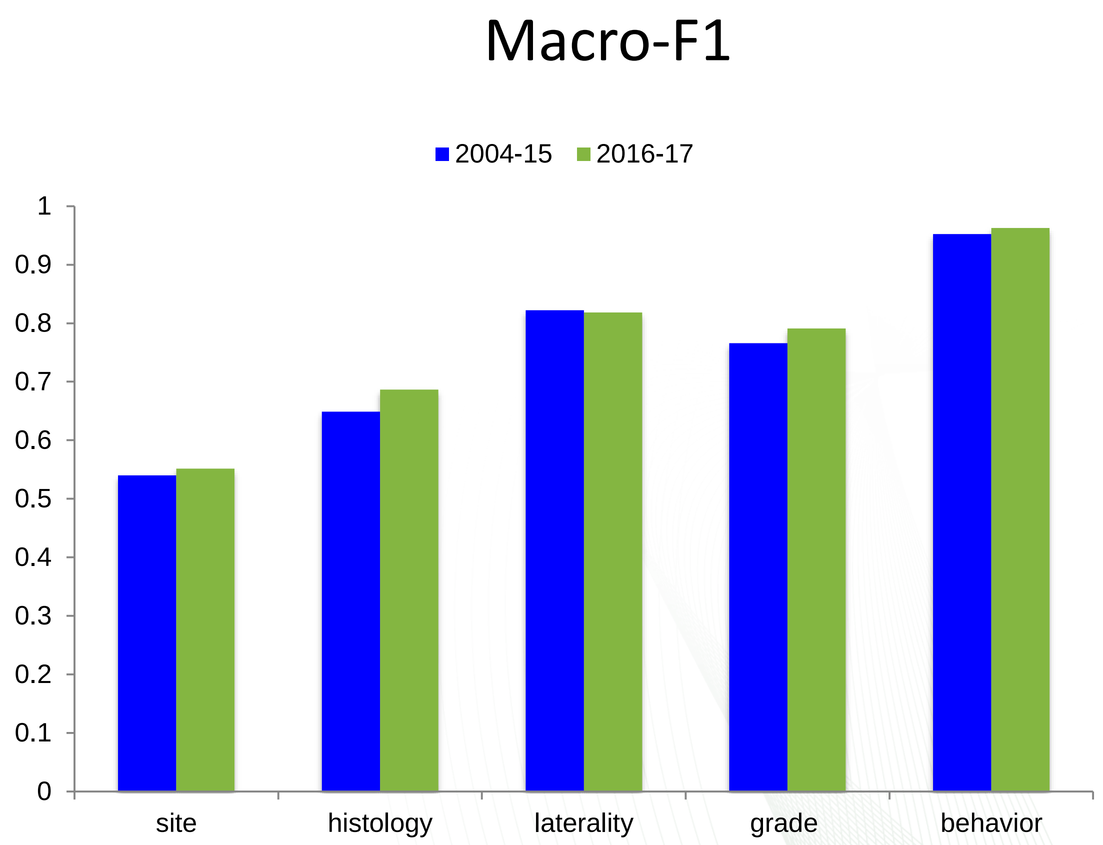
U.S. DEPARTMENT OF ENERGY | NIH NATIONAL CANCER INSTITUTE

Comparative Analysis: Multi-task CNNs perform better in information extraction tasks compared to single task CNN

	Single Task CNN		Multi-task CNN (Hard Parameter Sharing)	
Task	**Micro-F1**	**Macro-F1**	**Micro-F1**	**Macro-F1**
Site	0.8874	0.3643	**0.9401**	**0.5401**
Laterality	0.9079	0.6814	**0.9333**	**0.8222**
Behavior	0.9469	0.8840	**0.9746**	**0.9521**
Histology	0.7353	0.3638	**0.8206**	**0.6488**
Grade	0.7508	0.6820	**0.8023**	**0.7657**

Additional testing on 2016-17 cases



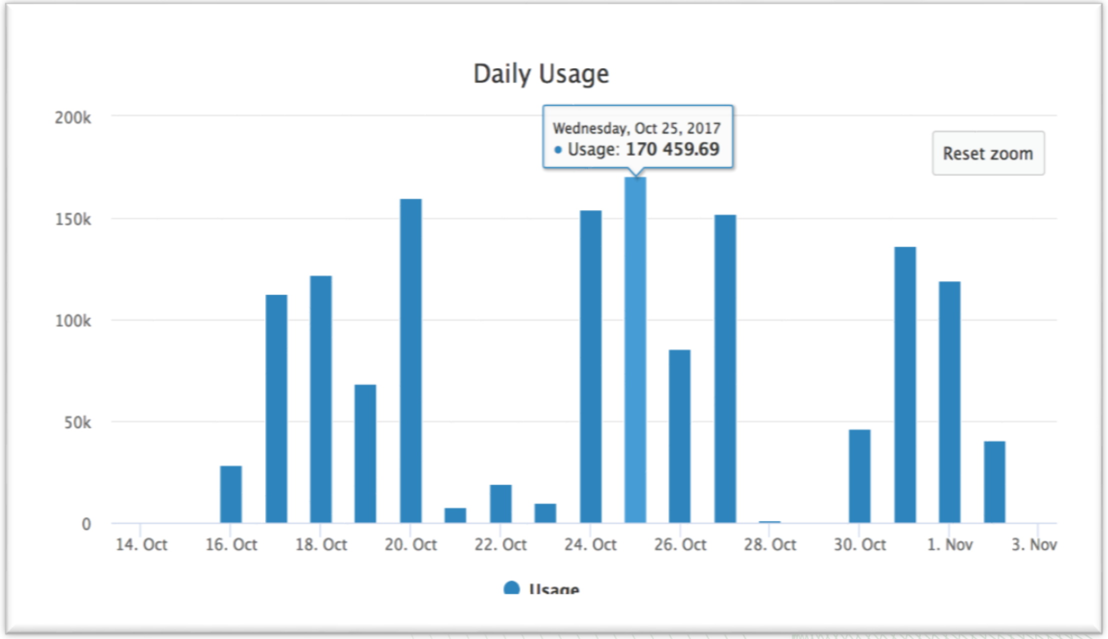


How fast can we train?

Experiments performed on OLCF infrastructure

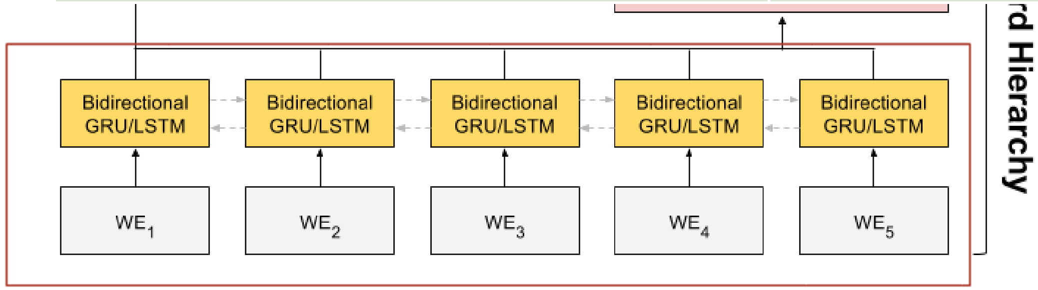
- **CNN Training on LA data**
  - 23,771 training cases
  - 5,942 validation cases
  - 29,714 testing cases
  - 50 epochs

	Titan	Summit
Platform Specs	18,688 nodes 1 x K20 GPU	4,600 nodes 6 x V100 GPU
Time	16.67 hrs	1.67 hours

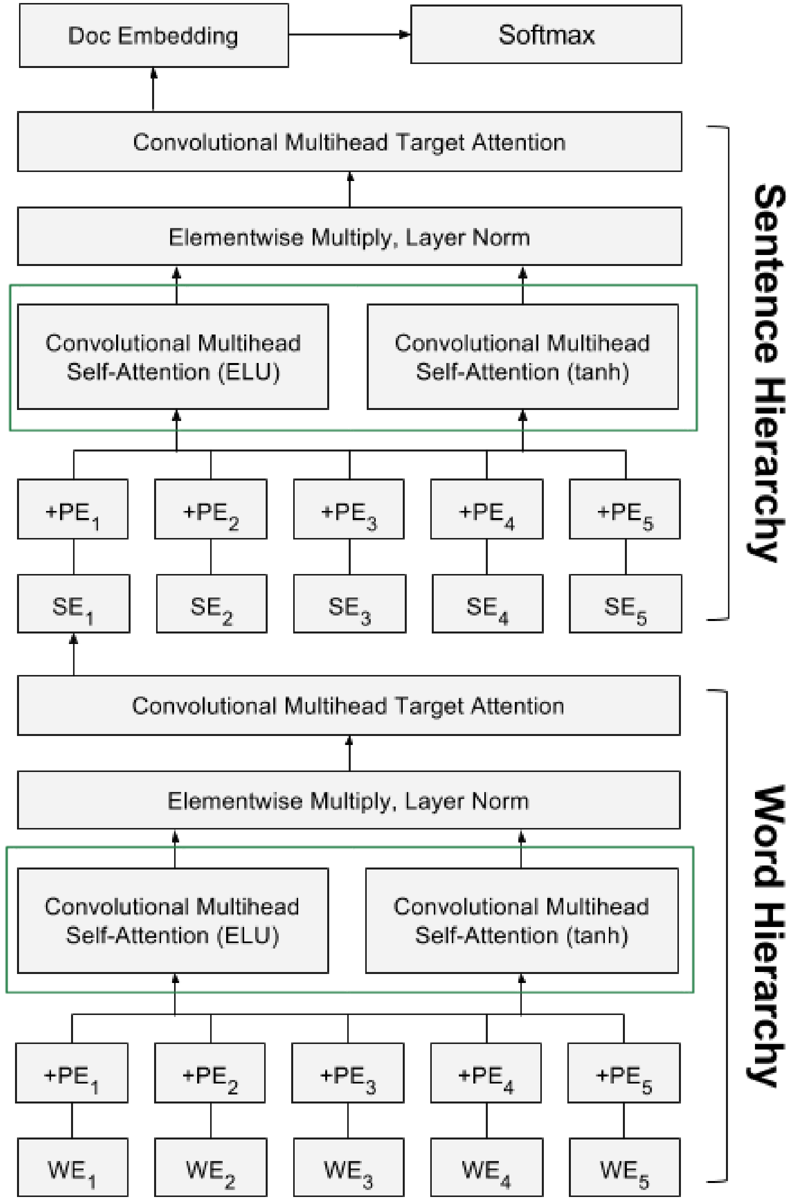


HAN is slow: Tweaking the network to accelerate training

	Pubmed
Naïve Bayes	76.63 --, 0.2s
Logistic Regression	76.46 --, 15s
CNN Baseline	77.25 13ms, 1hr
Hierarchical Attention Network	78.45 111ms, 9hr
Hierarchical Convolutional Attention Network	78.14 35ms, 3hr



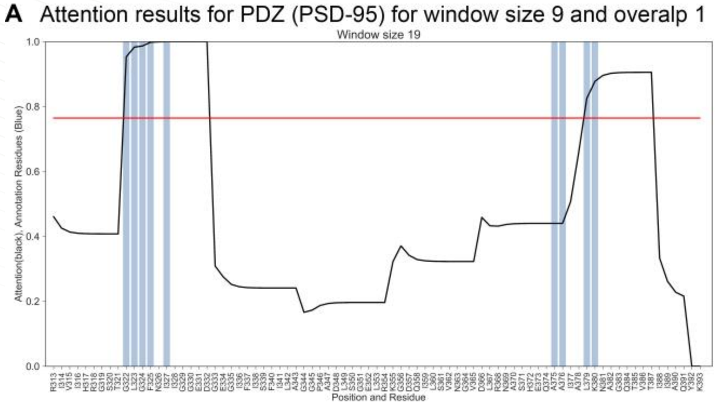
Computationally expensive!!!

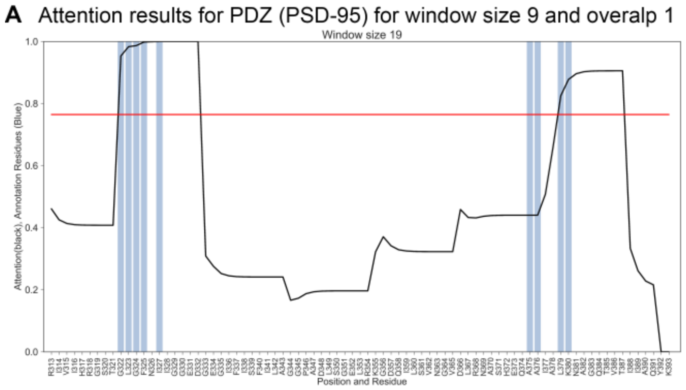


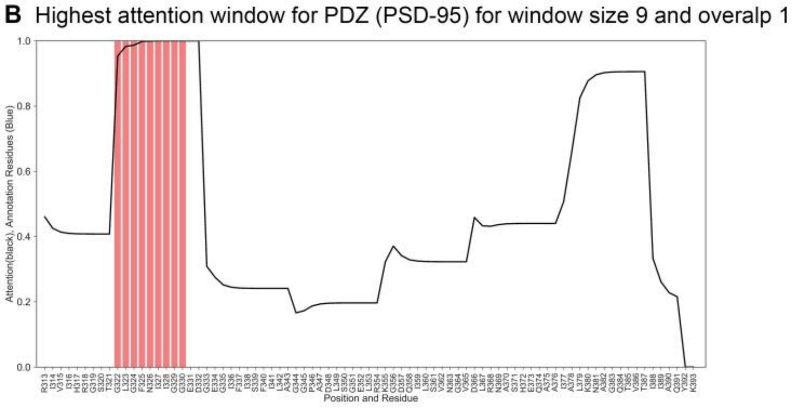
Gao, S., Ramanathan, A., in review (ACL)

Can the H(C)AN be used on other types of data? E.g., Protein alignments to understand co-evolutionary modules

- Predict “hotspots” across protein sequence databases







Protein Family	AUC (sequences)	F1 (sequences)	SCA AUC score	SCA F1 score
Cadherin	**0.568**	**0.817**	0.546	0.670
PDZ (NCBI)	**0.715**	**0.840**	0.520	0.753
PDZ (PFAM)	**0.660**	**0.827**	0.520	0.753
Tau	**0.555**	**0.643**	0.393	0.502
HSP70	0.510	**0.771**	**0.553**	0.709

Catanho, M., Gao, S., Ramanathan, A., Coleman, T. P., 2018 (submitted)

Summary & Conclusions

- CANDLE provides an enabling infrastructure for information extraction from clinical/pathology reports:
  - Simple DL networks provide good precision and sensitivity
  - Selection of DL networks is important to obtain good representations of data
  - Multi-task learning can exploit task relatedness and provide better results
- Development of semi-supervised learning approaches:
  - Lack of annotated text documents (labels)
  - Adversarial networks
- Predict the next “clinical state” of the patient from partial current clinical observations
  - Reinforcement learning/ Q-learning

Tomorrow's session

- Some lessons learned from working with CANDLE:
  - Preparing text (or related sequence) datasets for deep learning
  - Hyperparameter optimization
  - Any other questions regarding software or use

THANK YOU!!!

Questions/ Comments

ramanathana@ornl.gov